Caution: Rumors ahead—A case study on the debunking of false information on Twitter

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December: 1–15

DOI: 10.1177/2053951720980127
journals.sagepub.com/home/bds
SAGE

Caution: Rumors ahead—A case study on the debunking of false information on Twitter

Anna-Katharina Jung[1], Björn Ross[2] and Stefan Stieglitz[1]

Abstract

As false information may spread rapidly on social media, a profound understanding of how it can be debunked is required. This study offers empirical insights into the development of rumors after they are debunked, the various user groups who are involved in the process, and their network structures. As crisis situations are highly sensitive to the spread of rumors, Twitter posts from during the 2017 G20 summit are examined. Tweets regarding five rumors that were debunked during this event were manually coded into the following categories: rumor, debunking message, uncertainty about rumor, uncertainty about debunking message, and others. Our findings show that rumors which are debunked early and vehemently by official sources are the most likely to be stopped. When individuals participate in the process, they typically do so by sharing uncommented media content, as opposed to contributing user-generated content. Depending on the conditions in which a rumor arises, different network structures can be found. Since some rumors are easier for individuals to verify than others, our results have implications for the priorities of journalists and official sources.

Keywords

Social media, false information, rumor, debunking, fake news

Introduction

The veracity of information online has been intensively debated. Ideological polarization and a decreasing trust in traditional media catalyze the spread of false information (Spohr, 2017). Not least the disinformation campaigns during the US presidential elections sparked an interest in the role of fake news (Vargo et al., 2018). Since social media have become a key information source for many people, examining their role in the dissemination of false information is crucial (Nielsen et al., 2019).

Social media platforms take different approaches to this problem. Facebook's current strategies are to undermine the economic incentives and to develop technical solutions to help users to make more informed decisions, for example by showing warning messages, relying on the work of fact-checking units (Clegg, 2020; Ross et al., 2018). While Twitter has always underlined that the company is not the "arbiter of truth", they recently introduced warning labels and even delete harmful content (Crowell, 2017; Roth and Pickles, 2020). Although platforms continuously improve their measures against the spread of false information, every individual user should also, to the best of their ability and knowledge, reflect on the truthfulness of the information they encounter. However, echo chambers in which alternative narratives spread, fueled by selective exposure, can complicate the evaluation of information (Spohr, 2017; Starbird, 2017).

When an imbalance arises between the demand for information and the amount of factual information available, rumors begin to spread (Oh et al., 2013). Although rumoring is a typical human reaction to make sense of an event, this moment of speculation

[1]Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Duisburg, Germany
[2]School of Informatics, The University of Edinburgh, Edinburgh, UK

Corresponding author:
Björn Ross, School of Informatics, The University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom.
Email: b.ross@ed.ac.uk

leaves a lot of room for the spread of false rumors, or even mis- and disinformation, which can be dangerous (Mirbabaie and Marx, 2020).

A number of terms and concepts are used to describe and theorize false information. Fake news, mis- and disinformation, and rumors are the most prominent examples. In this study, we use the term rumor, as it is the broadest of these concepts. We understand rumors as "unverified and instrumentally relevant information statements in circulation that arise in the context of ambiguity, danger or potential threat, and that function to help people make sense and manage risk" (Bordia and DiFonzo, 2007: 13). The term is applied to a piece of information whose veracity at the time of dissemination is unclear; the possibility remains that it could turn out to be true (Spiro et al., 2012).

The desire for facts is especially high in moments of uncertainty and during opinion formation processes. Thus, the identification and rebuttal of false rumors is important for crisis management (Zeng et al., 2016). Due to the high level of uncertainty and the information need of the people involved, crisis situations are vulnerable to the spread of rumors (Oh et al., 2013). Although rumoring itself is a necessary process, the spread of false information can lead to panic reactions and false accusations and it might even constrain the work of emergency agencies (Abdullah et al., 2015). Thus, the dissemination of false rumors can be a security threat. Therefore, it is important that useful and factual information is disseminated to distinguish false from true rumors and to guide the collective sense-making process (Simon et al., 2016).

As even the most sophisticated detection algorithms will not entirely prevent the spread of false information, a profound understanding of how it can be corrected effectively is required. Only few studies have investigated the development of rumors during and after their rebuttal. Existing studies indicate that the total number of debunking messages is lower than the number of rumor-related messages, which implies a lower total reach (Chua et al., 2016; Starbird et al., 2014). Furthermore, the spread of rumors often continues even after they have been debunked, which is referred to as the "echo effect" (Jong and Dückers, 2016: 340). The prolonged spread of false rumors, after their correction, can be understood as a further indicator of an ineffective debunking strategy. We therefore address the following research question:

> RQ1 How does the spread of rumors develop after they are debunked?

A closer look at the users involved shows that the amount of discussion between the spreaders of rumors and corrections is limited (Bessi et al., 2015; Starbird et al., 2014). This indicates that filter bubbles can influence the effectiveness of debunking strategies. Official sources have been identified as important actors for rumor corrections, especially in crisis situations (Andrews et al., 2016). As official sources such as media organizations or emergency agencies are often in central network positions and enjoy a high level of credibility, the role of individual users in the debunking process has received less scientific attention. However, in times of increasing media bias and in societies with less press freedom, the influence of individuals can be important (Haigh et al., 2018). Their impact should not be underestimated. We therefore address the following research question:

> RQ2 How do individual users participate in rumoring and debunking processes?

To find generalizable patterns linked to the diffusion of rumors in social media, the network structures of five rumors are compared. The diffusion patterns of rumors on a network level are only sparsely researched (Bagavathi and Krishnan, 2019; Garcia-Herranz et al., 2014). The major focus of previous research often lies with rumors, ignoring the important debunking messages. To find out if different rumor types entail different network structures and thus might require different debunking strategies, we address the following research question:

> RQ3 Which network structures can be found in rumoring and debunking processes?

This article contributes to the literature a systematic comparison of the debunking of five false rumors for a political event in Germany, with a particular focus on the role of individuals and on network structures. 736,577 Twitter posts about the Group of Twenty (G20) summit in July 2017 in Hamburg were analyzed. We find that the rumors that were debunked vehemently and early were the most likely to be stopped. The professional media and official sources such as the police play a paramount role in this regard. Individual users then forward their statements to followers, which helps the debunking message spread. The networks of the rumor spreaders and debunkers can be distinguished into three distinct types, reflecting different underlying community structures, for example based on the users' different political views. Our work implies that journalists and authorities should focus on rumors that are hard to verify for people without privileged access to information. Rumors that are easy to recognize as false to anyone who is on site are more likely to self-correct.

Related work

False information on social media

The mechanisms of news broadcasting have changed fundamentally in the past decades. Digital media have altered our communication behavior (Newman et al., 2016). The rapid flow of information in social media facilitates the spread of fake news and unverified information, both for commercial and for political purposes (Chen et al., 2013). Virally spread false information can lead to panic reactions during crisis situations that cause economic and reputational damage (Chen et al., 2013; Oh et al., 2013; Wang et al., 2016).

There are numerous terms related to false information in social media: conspiracy theories (Bessi, 2017), false flags (Starbird, 2017), misinformation (Mazer et al., 2015), disinformation (Gupta et al., 2014), fake news (Rubin et al., 2015), satire (Cornwell et al., 2016), and rumors (Kwon et al., 2013). Many of these terms are used more or less interchangeably in academic literature, which underlines that their definitions are not clearly delineated (Al-Mansour et al., 2014; Rubin et al., 2015).

The most prominent concepts are mis- and disinformation, fake news, and rumors. We avoid the term fake news due to its politicized nature, especially since the 2016 U.S. presidential elections (Vosoughi et al., 2018). Disinformation is "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" (HLEG on Fake News and Disinformation, 2018). Misinformation does not imply intention (Hameleers et al., 2020) and it is therefore very close in meaning to rumors, but it is applied to information that is known to be false and it is therefore more negatively connotated. We decided to use the term rumors for this case study as it does not imply a judgment about the sender's intention or the veracity of the presented information.

Most approaches to identify rumors in social media rely on identifying deceptive content. Automatic identification techniques use indicators such as linguistic cues (Kwon et al., 2013), metadata about the sender (e.g. friend–follower ratio or verification status of Twitter accounts) and the presence of citations of external sources to back up a claim (Castillo et al., 2012; Gupta et al., 2014). As rumors especially arise in periods of high uncertainty, linguistic cues that indicate uncertainty have been used as early warning systems for rumors (Zhao et al., 2015). Another scientific aim is the development of automatic approaches to identify false information in real-time to prevent their diffusion. Most real-time solutions take the form of browser plugins, which warn the users about potentially false information (Chakraborty et al., 2016; Gupta et al., 2014). Although researchers have made progress in such automatic detection and flagging of rumors, it is unlikely that social media will ever be entirely free of misinformation.

Debunking false information on social media

The term *debunk* was coined by Woodward (1923) as a synonym for "take the bunk out of things". Once a false rumor spreads on social media, approaches are needed to correct it and reach those who have seen, liked or shared it. Jong and Dückers (2016) describe three ways to correct false information on Twitter: deleting the post, posting a correction, or directly contacting the rumor spreaders via @-mentions. Only the deletion of a tweet also deletes follow-up communication in the form of retweets. Copies of original rumor posts, e.g., in the form of screenshots, are problematic on Twitter and Facebook, as they can have an impact on the persistence of false information (Friggeri et al., 2014).

To understand the debunking process, it is essential to study the messages and their authors. Journalists and researchers are often thought to be responsible for the identification and correction of false rumors online (Berghel, 2017; Dale, 2017). While political rumors allow journalists time to check and research, rumors in crisis situations require a quick rebuttal to prevent panic situations (Andrews et al., 2016). Government organizations also need to intervene quickly with correct information through multiple channels, to prevent further damage (Oh et al., 2013). Social media accounts of official bodies such as media organizations, emergency agencies, or politicians have a high reach. In moments of uncertainty, social media users consult them, as they are considered reliable (Andrews et al., 2016). In moments of information scarcity, before official accounts are able to deliver trustworthy information, social media users turn to unofficial accounts (Oh et al., 2013; Shklovski et al., 2008). Andrews et al. (2016), who investigated the role of official accounts in the context of rumor propagation, showed that debunking messages posted by an official account motivate other users to correct earlier false information. According to Hunt et al. (2020), most of the debunking tweets by individual users are related to news agencies and government organizations. However, the findings also showed that messages affirming the rumor received more attention than those debunking it.

Journalists are thought to be in a position of great responsibility concerning the containment and correction of rumors due to their professional ethics (Arif et al., 2017). If they pick up breaking news from

social media without verification, they can foster the spread of rumors and strengthen their negative impact due to their reach. Jong and Dückers (2016) hypothesize that the media and official sources influence the rebuttal of rumors when this requires deeper investigations, such as getting in touch with officials, companies, or institutions, as this fact-checking can hardly be done without a press card or police ID. It needs to be doubted that social media users can always accurately identify trusted sources. Andrews et al. (2016) discovered the great influence of "breaking news accounts" in rumor propagation. These accounts imitate the layout and communication patterns of trusted media outlets. Although they were not real official sources, their tweets were highly shared.

A frequently presented hypothesis is that the crowd itself can detect, question, and even remove false information (Heverin and Zach, 2012). This assumption is linked to the concepts of collective intelligence and the wisdom of the crowds (Surowiecki, 2005), but it needs to be questioned. Studies show that misinformation on social media outnumbers debunking messages (Chua et al., 2016; Starbird et al., 2014). Although these findings cannot reflect whether corrections were seen and understood by many, they make it seem unlikely that effective self-cleaning takes place. Even rumors which have already been debunked are still shared, a notion known as the "echo effect" (Andrews et al., 2016; Chua et al., 2016; Jong and Dückers, 2016). In contrast to individuals, official sources are more cautious to share information that is already obsolete (Jong and Dückers, 2016).

While the motivation to share rumors has been investigated, there are few studies about motivations to correct them. Oh et al. (2013) showed that source ambiguity, personal involvement, and anxiety motivate users to share rumors online. Their findings underlined that source ambiguity plays less of a role on Twitter than it does in offline contexts. In contrast to rumors, corrections show a lower level of anxiety and emotions and are often factual and unambiguous (Chua et al., 2016). Regarding individual users, Arif et al. (2017) found that most users who affirmed a rumor did not take any corrective action because of continued uncertainty about the rumor and the feeling that there was no need to correct it. Additionally, Wang and Zhuang (2018) investigated the misinformation and debunking response of misinformed Twitter users, which supports the finding of Arif et al. (2017) that misinformed Twitter users often do not take any action after seeing a debunking message.

There has also been very limited research on the structures of the communities involved in sharing and debunking rumors. Social network analysis has become a widely accepted tool (Stieglitz et al., 2018), yet its application to rumor research is rare. There have been calls from scholars to use more tools from social network analysis, for example from the marketing literature, where the goal is to understand the characteristics of negative word of mouth (Pfeffer et al., 2014). In a recent study of online rumors, networks were used to visualize the relative frequencies of accurate and inaccurate tweets (Zubiaga et al., 2016). Characteristics of the network are sometimes leveraged to help improve the accuracy of rumor detection tools (Kwon et al., 2017; Zubiaga et al., 2018), but in these studies, the networks themselves are not studied to understand how rumors spread, or how they can be debunked effectively.

Materials and methods

Case description

On 7 and 8 July 2017, the heads of state of the world's largest economies met in Hamburg to discuss the global political situation at the G20 summit. Anti-globalization movements criticized this, and especially the demonstration "Welcome to Hell" on 6 July triggered violence. This resulted in the use of water cannons and tear gas by the police, to prevent vandals from torching cars and barricades or plundering nearby shops. Various rumors spread on social media. For this study, the five rumors were chosen that received the greatest attention online and offline and were debunked by credible sources while the event was still ongoing, namely the Hamburg Police Department and the fact-checking unit of the *Tagesschau*, the most popular TV news broadcast in Germany. This allows the analysis of the rumor and debunking messages over time. The *Tagesschau* is produced by Germany's public service broadcaster and it is one of the most objective German news sources, as it is neither state media nor a private TV station but financed by the German citizens. According to the Reuter's Digital News Report 2019, *Tagesschau* is the most trusted news brand in Germany (Nielsen et al., 2019). The rumors concern five alleged events, namely

1. the use of army tanks by the police against the protesters (*tank*),
2. a police officer being permanently blinded by a firecracker (*firecracker*),
3. a police raid on the left-wing venue *Rote Flora* (*RoteFlora*),
4. the user of nuclear weapons by the police to deter the protesters (*nuclear*), and
5. the protesters attacking a hospital and emergency room (*hospital*).

Since the focus of the study is on debunking, all the selected rumors were false, and debunked while the G20 summit was still underway.

Data collection

We focus on the microblogging service Twitter. In contrast to other social networks such as Facebook and Instagram, communication on Twitter is almost entirely public. The process of information diffusion on Twitter can be traced by the analysis of original tweets, retweets, commented retweets, replies, and likes of tweets. The use of Twitter by official institutions and media outlets to spread breaking news makes it useful for research questions related to communication during events and crises (Stieglitz et al., 2017).

Tweets in German published between 5 and 10 July 2017 00:00 UTC were collected using the Twitter Search API if they matched at least one of the following predefined keywords: *g20*, *nog20*, *g20-gipfel*, *g20summit*, *welcometohell*, *blockg20*, *g20ham17*. The selection of keywords for the tracking of Twitter data is an established approach. It leads to a reduction of noise in the data set and facilitates data cleaning. Neutral keywords (e.g. *g20* or *g20summit*) were used alongside those used by opponents of the summit (*nog20*, *welcometohell*). The data set includes all tweets of the accounts of the police department, the fire brigade, and the city of Hamburg, as those accounts are likely to be involved in debunking.

Qualitative identification of rumors and debunking messages

Tweets related to the selected rumors were identified. The total data set contains 736,577 tweets: 168,799 original tweets, the rest retweets. To identify the five rumors, a list of keywords was used (see Table 1).

After identifying the related tweets, a codebook was developed to categorize them (see Table 2). As rumors often spread in moments of ambiguity, the categories *uncertainty about rumor* and *uncertainty about debunking message* were added (Zeng et al., 2016; Zhao et al., 2015). All codes are mutually exclusive.

To meet reliability standards, two independent coders coded the data set, and Cohen's Kappa (κ) was calculated. The overall interrater agreement for the five rumors was 0.97, and between 0.92 for the *nuclear* rumor and 0.99 for the *hospital* and *firecracker* rumors. This level of reliability can be considered almost perfect (Landis and Koch, 1977). When coders one and two disagreed, a third coder was involved, and the majority rule was applied. To be able to answer RQ1, the development of the rumors and debunking messages were analyzed over time. The peaks of rumors and debunking messages were visualized and the impact of the debunking message on the dissemination of the rumor evaluated.

Identification of account roles and debunking message types

To analyze RQ2 and examine how various actors are involved in debunking rumors, the communication roles are classified. Mirbabaie et al. (2014) identified five primary roles: emergency services agencies, media organizations (including journalists and bloggers), political groups and unions, individuals (politically

Table 1. Filtering keywords and selection criteria for tweets (translated into English from German).

	Keywords	Selection criteria
Tank rumor	Tank, German armed forces	Selected: tweets stating or denying that tanks were seen or mobilized in Hamburg, that the German armed forces were in action, that a state of emergency had been declared; pictures of tanks in connection with the event Excluded: calls for the German armed forces to be deployed, reference to other kinds of armored vehicles, e.g., the armored vehicles of the police
Firecracker rumor	Blind, retina, firecracker, neck	Selected: tweets stating or denying that a police officer had been permanently blinded Excluded: metaphorical usage of the word "blind", such as the police being blind to something
RoteFlora rumor	Flora, raid	Selected: tweets stating or denying that the police had raided the left-wing venue *Rote Flora* Excluded: tweets that merely asked for the venue to be raided by the police or army
Nuclear rumor	Nuclear missile, rocket, bomb	Selected: tweets stating or denying that nuclear missiles were used in Hamburg by the police Excluded: tweets talking about an anti-nuclear treaty
Hospital rumor	St. Georg, emergency room, hospital	Selected: tweets stating or denying that left-wing extremists had attacked the hospital of St. Georg Excluded: tweets talking about police officers or protesters under medical treatment

Table 2. Coding scheme for rumor-related messages on Twitter (translated into English from German).

Category	Description	Code	Example
Rumor message	False information is forwarded. No doubt is expressed.	1	The emergency room of a hospital has been attacked, patients need to be evacuated, reports NTV #schanze #welcometohell #G20HH2017.
Debunking message	A rumor is denied or a correction is shared. The rumor is corrected in the tweet itself or a linked article.	2	Again, for EVERYONE! There is no nuclear missile in use at #G20.
Uncertainty about rumor	The tweet shares the rumor, but questions it, possibly involving credible sources such as the police or army via @-mentions.	3	Are the #Germanarmedforces coming for further reinforcement? (The police have not confirmed that so far!)
Uncertainty about debunking message	The tweet shares a message debunking the rumor, but questions this message, possibly involving credible sources such as the police or army via @-mentions.	4	G20 summit Hamburg: Military vehicles in Hamburg—allegedly a parking problem.
Others	Jokes, unclear statements, and opinions.	5	Why the left-wing extremist #RoteFlora is not raided and shut down is something that only the left-wing politicians and judiciary understand.

engaged or personally involved), and commercial organizations. For this study, the categorization was slightly adapted. To improve the accuracy of the results, the group *media organizations* only includes official media accounts. Journalists and bloggers who use their personal accounts are coded as the subcategory *journalist (personal)*. To only consider media organizations would ignore the usage of Twitter by journalists and bloggers in their free time, although they post in a more professional manner than others. The category of commercial organizations was dismissed, as it was not applicable.

To identify accounts by individuals, which were of particular interest in this analysis, their verification status was considered. Twitter verifies accounts of public interest, such as those run by major news outlets, organizations, politicians, and celebrities. A badge is displayed on these accounts' Twitter pages.

Following these criteria, the roles of the 10 most retweeted users spreading debunking messages and rumors were determined. To ensure that the manual coding process met reliability standards, two independent coders coded the data set and Cohen's kappa (κ) was calculated. Interrater agreement was 86.8%. In cases of disagreement, a third coder was involved, and the majority rule was applied.

Furthermore, we analyzed which rumor tweets were deleted by the users or platform, to understand the deletion patterns. This was done by a manual check of the tweet ID after the final data collection. Finally, the content of the debunking messages was examined to identify debunking strategies.

Social network analysis

To examine the involvement of the different users, their positions in the network and their influence on the rumoring and debunking process (RQ3), the social networks formed by the participating users were visualized. A retweet network was created for each rumor from the tweets that mention the respective rumor. In such a network, each Twitter account is a node. A directed edge from node A to node B indicates that A retweeted a tweet that was originally published by B.

The network was visualized with the open-source software Gephi (Bastian et al., 2009). Nodes were colored according to the type of messages that they had spread regarding the rumor: red for accounts that spread rumor messages, blue for debunking messages, and white for users who had spread messages of both types. For simplicity, nodes were not included if they only spread tweets in one of the "uncertainty" categories or "other". The size of the nodes corresponds to their in-degree, that is, the number of accounts that retweeted their tweets, as a measure of the influence of an account. The algorithm Force Atlas 2 was used to calculate the graph layout (Jacomy et al., 2014). In this physics-inspired algorithm, nodes repulse each

other, and edges attract their nodes, creating a map in which densely connected subgroups of users are close together and sparsely connected groups far apart.

Results

Development of rumors and debunking messages over time

A total of 6095 tweets were identified that are related to one of the five rumors (Table 3). This number includes both original tweets and retweets. The rumor with the highest reach is the allegation that the police were planning to use nuclear weapons to deter the protesters (rumor *nuclear*). Almost all tweets about this rumor (95.1%) were coded as spreading the rumor, and very few (1.7%) as debunking messages. The level of uncertainty is low (2.4%). The most retweeted user for rumor *nuclear* is the satire magazine "Der Postillon", which published this satirical news story. The *tank* rumor is the subject of the second highest number of tweets. Almost half (47.6%) of the tweets about it spread the rumor and a third (33.1%) were debunking messages. The *tank* rumor shows the highest level of uncertainty: more than one in six tweets (17.8%) express uncertainty about the rumor or debunking message.

While there are more rumor messages than debunking messages about the *tank* and *nuclear* rumors, the situation is the reverse for the *firecracker*, *RoteFlora*, and *hospital rumors*. The *firecracker* rumor is the subject of almost twice as many debunking messages (61.1%) as rumor messages (33.3%). Similarly, while more than a third (37.6%) of tweets about the *RoteFlora* rumor, the alleged raid on the left-wing venue, help spread it, more than half (51.2%) contribute to debunking it. While more than 10% of the tweets show uncertainty about the rumor, there is no uncertainty about the debunking claims. The *hospital* rumor shows the highest share of debunking messages (88.4%). Less than one in ten (9.8%) spread the rumor without at least expressing some uncertainty about it.

The five selected rumors behave differently over time (Figure 1). As typical for rumors in crisis situations, each rumor peaks and then quickly subsides. The *tank* rumor shows a concentrated peak of rumor tweets between 10 a.m. UTC and 12 a.m. UTC on 7 July. The first debunking messages spread during the main rise of the rumor at 11 a.m. and reach their peak at 12 a.m. Most of the debunking messages are retweets of the tweet by the police department. At 12 a.m., the dissemination of the rumor and debunking messages starts to decline, at 2 p.m. the rumor already drops to less than 50 tweets per hour. At midnight, the rumor fully loses the users' attention. Between its first appearance and midnight, claims expressing uncertainty about the rumor and debunking messages circulate. The satirical *nuclear* rumor also peaks at 11 a.m. on 7 July, with 458 tweets per hour.

In contrast to the *tank* and *nuclear* rumor, the conversation around the *firecracker*, *RoteFlora*, and *hospital* rumors is dominated by debunking messages. The *firecracker* rumor begins with a low but constant spread of misinformation. On 8 July at 11 a.m., there is a minimal rise in rumor messages up to 23 tweets per hour. At around the same time, the first debunking messages appear and are retweeted. This leads to a drop in the number of rumor messages. The rumor is last shared on 9 July at 12 a.m. On its peak on 8 July at 2 p.m., the retweet of the official rebuttal by the police department and a tweet by the editor-in-chief of the newspaper BILD Hamburg fuel the spread of debunking messages.

The *RoteFlora* and *hospital* rumors develop similarly over time. The *RoteFlora* rumor peaks on 7 July at 7 p.m., due especially to retweets of comments by active individuals who spread the rumor that the

Table 3. Distribution of original tweets and retweets of all codes by rumor ($n = 6095$).

Code	Tank rumor Use of tanks by police	Firecracker rumor Police officer permanently blinded	*RoteFlora* rumor Police raid on left-wing venue	Nuclear rumor Use of nuclear weapons by police	Hospital rumor Attack on emergency room by protesters
Rumor	840	279	181	2335	54
Debunking message	584	513	247	41	488
Uncertainty about rumor	154	43	51	52	10
Uncertainty about debunking message	161	2	0	7	0
Other	27	2	3	21	0
Total	1766	839	482	2456	552

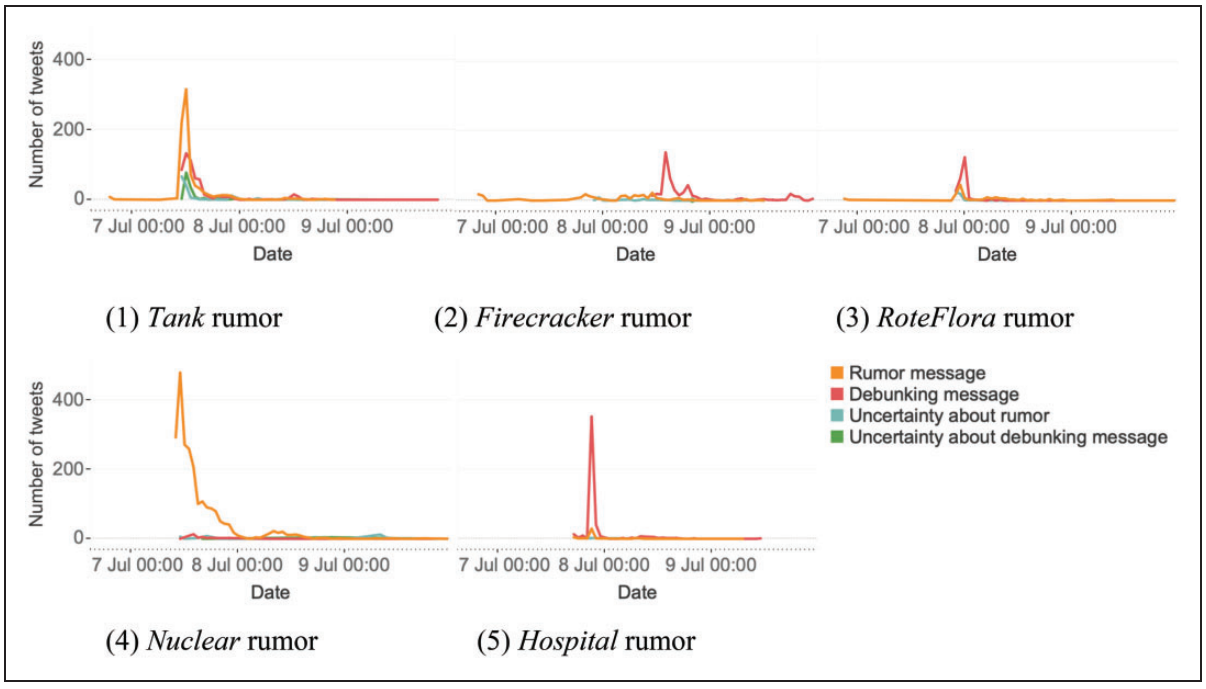

Figure 1. Development of rumor messages, debunking messages, and uncertainty over time. 1: tank, 2: firecracker, 3: Rote Flora, 4: nuclear, 5: hospital rumor.

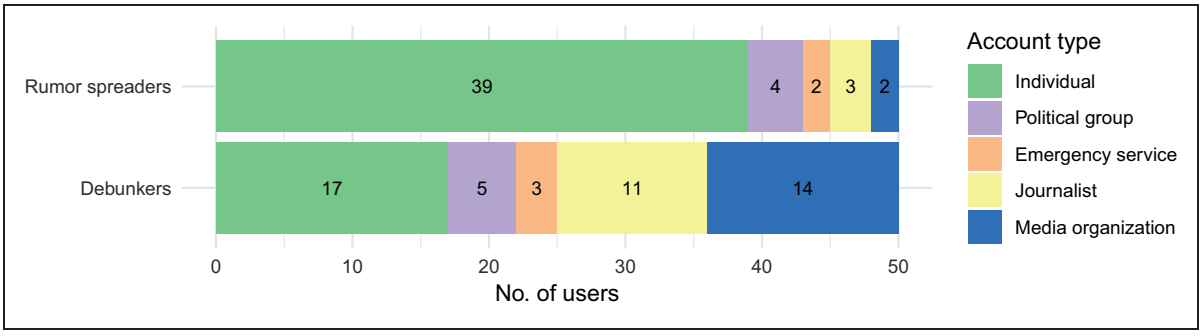

Figure 2. The top 50 rumor spreaders (the 10 most retweeted rumor-spreading accounts for each rumor) and the top 50 debunkers by type of account. Most of the most widely retweeted rumor spreaders were individual (unverified) users. Individuals also play a large role in debunking the rumors, but so do (verified) professional journalists and media organizations.

left-wing venue Rote Flora had been raided. The number of debunking messages rises simultaneously with the rumor messages and peaks at midnight. After this, the rumor quickly vanishes. The debunking messages are mostly retweets of two left-wing groups. The *hospital* rumor starts at 5 p.m. with tweets by two unverified user accounts who share that a police officer they know had confirmed an attack on a hospital. Quickly, the first debunking messages are shared by the same left-wing group that was also involved in debunking the *RoteFlora* rumor. On 7 July at 9 p.m., the debunking messages peak, including official debunking messages by the police department and the public service broadcaster NDR. The spreading of the rumor drops at the same time as the debunking messages peak. At around 11 p.m. on 7 July, the rumoring process is over.

Participation of individuals in debunking rumors

To understand how rumors can be debunked effectively on social media, it is necessary to study the roles of the users involved in the communication and the types of messages they post to debunk rumors. Figure 2 shows the top users involved. It shows that the largest group, both among the rumor spreaders and among the

debunkers, is that of individual (unverified) users. Almost four in five of the top 50 rumor spreaders are individual users, as are almost two in five of the top debunkers.

The involvement of individual users merits a closer look at their debunking messages. The most common type is the uncommented retweet of a rumor correction by an official source. The most active such "official" debunker is the Hamburg police department, with eight debunking messages and 873 retweets. For the *tank*, *firecracker*, *nuclear*, and *hospital* rumor, it is among the top five most retweeted accounts. The media and journalists are also highly represented in the group of official debunkers.

In contrast to retweets of debunking tweets by others, there are also original debunking tweets by individual users. In the tweets by these users, three strategies can be identified: sharing media content or official statements by emergency services with one's followers that deny a rumor, addressing the rumor spreader directly by @-mentioning them, or making an unsourced claim to one's followers that a rumor is false. Figure 3 shows how common each of the strategies was for each of the five rumors. Overall, the most popular strategy is to share a link to a statement by the emergency services or professional journalists explaining that the rumor is false.

Finally, an approach that contributes to stopping the rumor spread, although not to debunking it, is to delete the tweet that contains the false claim. Of the tweets about the *tank* rumor that were identified as rumor messages (i.e. contributed to spreading it), 13 original tweets (13.2%) were deleted. Of the tweets spreading the *firecracker* rumor, 33% were later deleted. This is mainly linked to the suspension of a highly active unverified user account, who wrote four tweets that were shared by 82 individual users. Rumor *RoteFlora* has a low deletion rate of rumor messages (4.4%). They cannot be accessed any longer as the original account is suspended. It is the same individual user who was responsible for the most deletions of messages spreading the *firecracker* rumor. The satirical *nuclear* rumor has the lowest deletion rate (1.2%). Of the tweets spreading the *hospital* rumor, 37% were later deleted. The analysis of the users who deleted their rumor tweets shows that only unverified accounts of individuals used this approach.

Rumor and debunking networks

Finally, we examine the network structures that shape rumoring and debunking processes. The network visualizations (Figure 4) reveal that different rumors can have vastly different community structures. They can be roughly classified into three categories: mixed networks, polarized networks, and one-sided networks.

The mixed networks of the *tank* and *firecracker* rumors are characterized by a relatively high number of ties between those sharing rumor-related messages and those sharing debunking messages. This indicates that the same Twitter profiles that helped spread the rumor also retweeted messages by the Twitter profiles that helped contain or debunk it. This can be underlined, for example, by the account of the Hamburg police, which posted the most relevant debunking message, but also posted a tweet that contributed to the spread of the main rumor. One of the most relevant

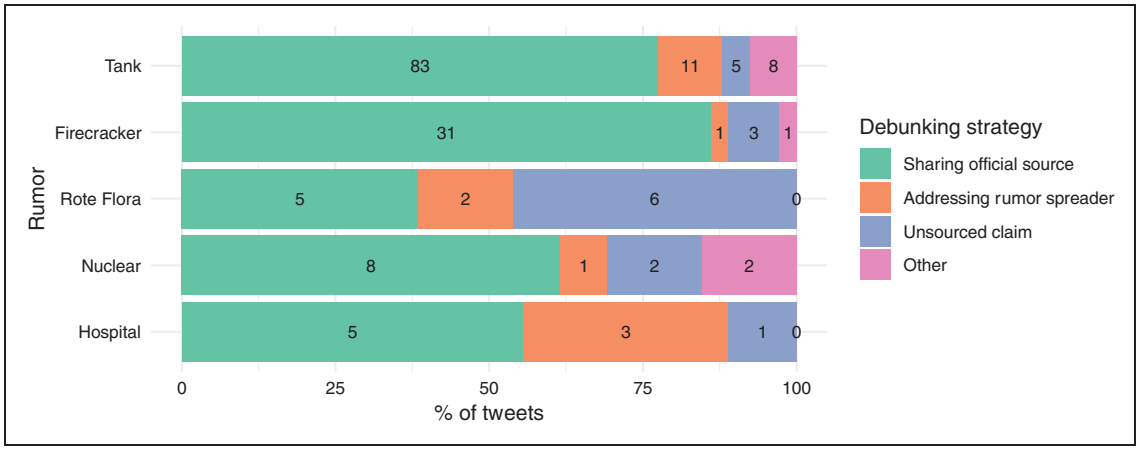

Figure 3. The strategies used by individual users to debunk the five rumors, excluding uncommented retweets. 1: tank, 2: firecracker, 3: Rote Flora, 4: nuclear, 5: hospital rumor. The figure includes all original debunking tweets by unverified accounts ($n = 178$). The lengths of the bars indicate the relative share of the strategy, while the numbers indicate the absolute number of tweets. The most popular strategy was to link to an official source disputing the rumor's veracity, such as an emergency service or a media organization. However, in the *RoteFlora* rumor, claims of the rumor being false were more likely to be made without providing a source.

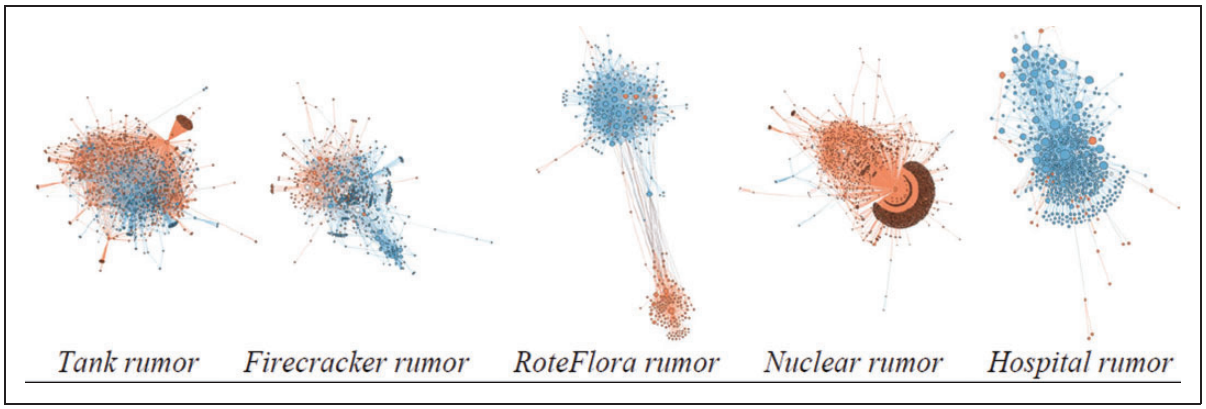

Figure 4. Networks of rumor sharers (red), debunkers (blue), and those who spread messages of both types (white). The force-directed drawing algorithm causes structural communities to form visible clusters (Jacomy et al., 2014). In the mixed networks 1 and 2, rumor sharers and debunkers are highly interconnected. The network for the *RoteFlora* rumor is polarized. The nuclear and hospital rumors are examples of one-sided networks, where individuals from the minority group hardly play a role.

accounts for the *tank* rumor is the individual unverified user who first shared the picture of the tank. This original post was heavily retweeted and thus this user has a central position among the rumor spreaders. This user can be found in the top right corner of the network visualization.

For *RoteFlora*, the situation is very different. The network is polarized. The force-directed drawing algorithm, which is based on the ties between the nodes, reveals two separate groups. There are far fewer ties between these groups than within each group. Of course, since critics of the G20 summit are left-wing activists, the entire discussion is political in nature, so some degree of polarization is to be expected. Yet, the difference between the two groups is much more pronounced than for the other rumors, suggesting that something else is at play here. We therefore inspected the network more closely. Most users in the smaller group, which consists entirely of rumor sharers, are far-*right* activists who expressed their enthusiasm at the news that the left-wing venue was being raided, with one user even calling for it to be set on fire.[1] Those in the larger group, which mostly consists of debunkers, are generally left-wing activists. There are only a few rumor sharers in this larger group, and they are also left-wing activists, such as one who, having heard of the impending raid, called for other protesters to meet at the venue.[2] Only few accounts link the two groups, because their debunking messages are retweeted by members of both groups. The one with the highest degree is the news station n-tv.

Finally, the *nuclear* and *hospital* rumors are examples of one-sided situations in which the number of those involved in spreading the rumor, or debunking it, respectively, is far greater than the number of accounts involved in the opposite activity, so that they dominate the network. For the *nuclear* rumor, "Der Postillon", the website that posted the satirical article about the alleged use of nuclear weapons by the police, is clearly visible in a very central role. The Hamburg police account is the most central one for the *hospital* rumor, as it posted the most retweeted tweet, a debunking message.

Discussion

This study offers insights into the spread and debunking of rumors on social media, in particular in the German Twittersphere. Our results should be interpreted in light of the fact that the study was conducted in a country where Twitter ranks among the five most important social networks but has a smaller and older user base than, for example, in the United States (Nielsen et al., 2019), and where trust in the news sector is much higher (Newman et al., 2016). The contribution of this article to the literature is the systematic comparison of the debunking of different rumors in the context of one political crisis event, with a particular focus on the role of individuals. Although the rumors took place over the same time span, they each behaved quite differently from one another. In the following, we highlight and discuss key differences that help explain why some of them were debunked more effectively than others.

The G20 data sets reveals unexpected patterns regarding the distribution of rumor messages, debunking messages and the levels of uncertainty: for three out of the five rumors, there were more debunking than rumor messages. This contrasts with previous findings,

where Twitter rumor messages outnumbered debunking messages (Chua et al., 2016; Starbird et al., 2014). Wang and Zhuang (2018) who analyzed rumor and debunking responses on Twitter for four rumors during Hurricane Sandy and the Boston Marathon bombing only identified one rumor which resulted in more debunking than rumor-related tweets. They hypothesized that the respective rumor was easier to debunk; however, they did not offer an explanation. Our result that for three out of five rumors there were more debunking than rumor messages underlines that there are different types of rumors, which differ in how hard they are to debunk.

The development of the rumor and debunking messages over time underlines that the reach of rumors in crisis situations depends greatly on temporal factors. This situation is very different from political rumors, which are more persistent (Abdullah et al., 2015). Although that might lead to the conclusion that the influence of rumor messages is limited, it needs to be stressed how dangerous the spread of false information in crisis situations can be. In this case study, for example, the spread of the *firecracker* rumor was linked to public calls to find the protester who threw the firecrackers that allegedly permanently blinded a police officer. In the middle of a heated protest, this situation can quickly become dangerous.

In the satirical news story alleging that the police were planning to use nuclear weapon, rumor tweets far outnumbered debunking messages. Only very few rumor tweets were later deleted. The results indicate that the Twitter community considered the rumor so absurd that they did not think it was necessary to debunk it. It can hardly be said that the satirical story was mistaken for real news, as Cornwell et al. (2016) anticipated. A self-correcting mechanism could be observed, as expected by Jong and Dückers (2016).

In debunking the *firecracker* rumor, an article by the newspaper BILD played a crucial role. However, some of the rumor tweets also refer to the same article. Although BILD later removed the incorrect claim from the article, they did not publish a separate tweet to inform users of the change. Therefore, it needs to be assumed that the users who shared the first version of the article did not realize that it was later corrected. While journalists should correct information publicly, as they greatly influence opinion formation, this ethical requirement collides with their fear of reputational damage. We argue that a tightening of journalism ethics and legal requirements might counteract this immoral behavior. However, a study on the US media landscape revealed that partisan media are highly involved in the production and spread of fake news, which is why not all media outlets might be interested in correcting false rumors (Vargo et al., 2018). Therefore, it might be interesting for platform providers to develop technical measures which identify and indicate updates to hyperlinked articles directly. The behavior of BILD in our case study complements the findings of Zubiaga et al. (2016), who observed that news organizations tend to use sources as an evidence in their tweets, but sometimes do not verify those sources and consequently contribute to the spread of rumors.

This paper provides results on which rumors are likely to be debunked early and effectively. Like Zubiaga et al. (2016), we recorded the greatest attention for rumors in their unverified stage. Especially the *RoteFlora* and *hospital* rumor show that the more vehemently and the more quickly rumors are rebutted, the more successfully they are contained. This confirms the findings by Oh et al. (2013) and Andrews et al. (2016). Additionally, the rumors circulating in the morning (*tank* and *nuclear* rumor) show a greater total reach than those spreading in the evening or at night (*RoteFlora* and *hospital*), which might be linked to the users' overall activity level. Moreover, the analysis shows that the rumor that was nearly impossible for individual Twitter users to disprove (the *tank* rumor, which would require them to contact the armed forces), showed the highest level of uncertainty and a low share of debunking messages. In contrast, the *RoteFlora* and *hospital* rumors, which could be disproved by anyone on site, were debunked quickly. The influence of users on site could also explain the quick resolution of one rumor in the case study of Wang and Zhuang (2018), for which they did not have an explanation. We conclude that timely and vehement debunking messages, not only by official sources, but also by users on site, are the most likely to stop the rumoring successfully. The more difficult it is for individuals to disprove a message, the higher the level of uncertainty and the greater the responsibility of official sources to debunk it.

This case study shows that although individual users form the biggest group of rumor spreaders (78% of the top 50), they also make up a third of the 50 top retweeted debunkers, which shows that they are very influential in this regard. However, as already shown in prior studies, the role of official accounts on debunking was high. The Hamburg police department was among the most retweeted debunkers for four of the five rumors, which clearly underlines that police departments need to train staff on how to detect and debunk rumors in social media. To be able to react promptly and determinedly to online rumors, debunking guidelines need to be developed that consider the different types of rumors. While our findings suggest that satire only requires a rebuttal if it is misinterpreted by many users, especially those rumors which are hard

to verify by the individuals should be quickly rebutted by the police. The impact of government accounts in debunking could also be found in the study of Hunt et al. (2020), who conducted a case study on the use of Twitter during two Hurricanes in the United States. The fact that the original debunking messages by the users mainly contain media content underlines that the media bear a great responsibility in debunking rumors, too. Although it is important that official sources are distributing facts during a crisis to debunk false rumors, this ideal has not been globally reached. The freedom of the press is at risk in many countries around the world and even in democratic societies, political leaders contribute to the spread of false information (Allcott and Gentzkow, 2017; Reporters Without Borders, 2020). Especially in those societies, individual debunkers and independent fact-checking institutions are important (Haigh et al., 2018).

Looking at the users' debunking behavior (RQ2), the main type of debunking message is the uncommented retweet of a tweet by an official or a professional journalistic source. Even when individual users write original tweets, these tweets are dominated by media content. This again underlines the importance and social responsibility of media institutions and official bodies. Both should be aware that they are in a position from which rumors can be effectively corrected. Only the *RoteFlora rumor*, which can be considered relatively easy to debunk by anyone who is on site, contains more unsourced claims that the rumor is untrue than media content. The *hospital* rumor shows a balanced picture, while in the *tank rumor*, which is more difficult to debunk, almost four in five debunking messages by individuals cite media content. A rarely used strategy to debunk rumors is to address the rumor spreader directly using an @-mention. The findings suggest that users hesitate to directly contradict others, and instead prefer to aim debunking statements at their own followers. This adds to the finding of Wang and Zhuang (2018) who also found that conversational debunking in the form of replies was a rarely used debunking method. To analyze the characteristics of a rumor and the level of difficulty to debunk it with the different debunking methods of individual users is a contribution of this case study.

Deleting rumor-related tweets is another approach to debunking rumors. In this study, it was only used by individuals. Yet, verified institutional Twitter accounts are also less likely than individuals to spread rumors in the first place, which makes the deletion of tweets less necessary. These results add to findings (Jong and Dückers, 2016) that official accounts are less prone to continue spreading rumors after they are debunked. The findings about rumor deletions show that the suspension of users by Twitter effectively reduces the number of rumor-related tweets. Thus, in the case of illegal behavior such as calls for violence, reporting the perpetrators can have a great impact, especially when the rumor spreaders have a high reach (see *firecracker* and *RoteFlora* rumors). Since the *tank*, *firecracker* and *hospital* rumors show a deletion rate higher than the average deletion rate of 11% (Bhattacharya and Ganguly, 2016), it can be assumed that tweet deletion is likely during and after the rumoring process.

This paper contributes to the literature by examining the networks of rumor sharers and debunkers (RQ3). In two of the networks, the users were mixed together in one large group. In another rumor network, however (*RoteFlora* rumor), the network is much more polarized. The final two rumor networks are very one-sided. The one-sidedness of these networks clearly mirrors the fact that the messages of one type were much more frequent than the messages of the other type. This community structure is a logical consequence of the numbers (see Table 3). In contrast, the difference between the *tank* and *firecracker* rumors and the *RoteFlora* rumor is not due to any difference in numbers. This phenomenon was only brought to light by the network analysis. In the *RoteFlora* rumor, the community structure mirrored the political views of the Twitter users. Those on the political right spread the rumor, those on the left the debunking messages. The most important connection between them was a news organization that spread a debunking message, once again highlighting the paramount role of professional journalism. However, this network structure might be an indicator of ideological echo chambers and a selective exposure of the users, who are only connected through few bridging accounts, as described by Spohr (2017). Only few previous studies have compared the retweet networks for the spread and debunking of multiple rumors. Zubiaga et al. (2016)'s approach has similarities to ours, although the circular layout chosen there does not allow an interpretation of the position of the nodes. The force-directed layout algorithm chosen in this work is suitable for clearly separating communities from one another if they have few connections in common.

It is still unclear why the *RoteFlora* network is so much more polarized than the networks for the *tank* and *firecracker* rumors. This result raises follow-up questions that should be addressed properly in separate studies: Which factors determine the emergence of mixed and polarized networks? How frequent are each of these network structures? Are rumors easier to debunk in networks with more interaction between the groups? The present findings show that network analysis is a viable tool that should be used in future research on the debunking of rumors.

This study has some limitations. The Twitter data set about the G20 summit only includes tweets which contained at least one of the selected keywords. The selection of five rumors can only give a limited overview about the rumors which circulated during the G20 summit. Furthermore, the manual coding of such data sets is never free from the coders' subjective interpretations, and sometimes ambiguous. As an example, a newspaper article that had at first helped spread the *firecracker* rumor was later corrected, so that those subsequently spreading it contributed to debunking the rumor without necessarily being aware of that.

Through the analysis, it has emerged that one of the key characteristics that distinguishes rumors that are debunked quickly from those that are not is the difficulty for individual users to research their veracity, without the privileged access to information that journalists and government agencies have. In the future, similar studies should be conducted in a different application area to test if this finding can be transferred to debunking in other contexts.

Acknowledgments

We would like to thank our student assistant Vera Vohwinkel for her dedicated support during the coding process of the G20 2017 Twitter data set.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 823866.

ORCID iDs

Anna-Katharina Jung https://orcid.org/0000-0002-0905-4932

Björn Ross https://orcid.org/0000-0003-2717-3705

Notes

1. This account has been suspended by Twitter.
2. This account has been suspended.

References

Abdullah N, Nishioka D, Tanaka Y, et al. (2015) User's action and decision making of retweet messages towards reducing misinformation spread during disaster. *Journal of Information Processing* 23(1): 31–40.

Al-Mansour A, Brankovic L and Iliopoulos CS (2014) A model for recalibrating credibility in different contexts and languages – a twitter case study. *International Journal of Digital Information and Wireless Communications* 4(1): 53–62.

Allcott H and Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2): 211–236.

Andrews CA, Fichet ES, Ding Y, et al. (2016) Keeping up with the tweet-dashians: The impact of 'official' accounts on online rumoring. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing – CSCW '16,* 2016, pp.451–464. New York, NY: Association for Computing Machinery.

Arif A, Robinson JJ, Stanek SA, et al. (2017) A closer look at the self-correcting crowd. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing – CSCW '17*, 2017, pp.155–168. New York, NY: Association for Computing Machinery.

Bagavathi A and Krishnan S (2019) *S*ocial sensors early detection of contagious outbreaks in social media. In: Ahram T (ed) *Advances in Artificial Intelligence, Software and Systems Engineering*. AHFE 2018. Advances in Intelligent Systems and Computing, vol 787, pp. 400–407. Cham: Springer.

Bastian M, Heymann S and Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. In: *Proceedings of the third international ICWSM conference 2009*, 2009, pp.361–362. Menlo Park, CA: The AAAI Press.

Berghel H (2017) Lies, damn lies, and fake news. *Computer* 50(2): 80–85.

Bessi A (2017) On the statistical properties of viral misinformation in online social media. *Physica A: Statistical Mechanics and Its Applications* 469: 459–470.

Bessi A, Coletto M, Davidescu GA, et al. (2015) Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE* 10(2): 1–17.

Bhattacharya P and Ganguly N (2016) Characterizing deleted tweets and their authors. In: *Proceedings of the tenth international AAAI conference on web and social media*, pp.547–550. Palo Alto, CA: AAAI Publications. Available at: www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13133

Bordia P and DiFonzo N (2007) *Rumor Psychology: Social and Organizational Approaches*. Washington, DC: American Psychological Association.

Castillo C, Poblete B and Mendoza M (2012) Predicting information credibility in time-sensitive social media. *Internet Research* 23(5): 528–543.

Chakraborty A, Paranjape B, Kakarla S, et al. (2016) Stop clickbaiting: Detecting and preventing clickbaits in online news media. In: *IEEE/ACM international conference on advances in social network analysis and mining (ASONAM)*, 2016, pp.9–16. New York: IEEE.

Chen C, Wu K, Srinivasan V, et al. (2013) Battling the internet water army. In: *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks*

analysis and mining – ASONAM '13, New York, USA, 2013, pp. 116–120. New York: ACM Press.

Chua AYK, Cheah S-M, Goh DH, et al. (2016) Collective rumor correction on the death hoax. In: *PACIS 2016 proceedings*, 2016. Available at: http://aisel.aisnet.org/pacis2016/178

Clegg N (2020) Combating COVID-19 misinformation across our apps. Available at: https://about.fb.com/news/2020/03/combating-covid-19-misinformation/ (accessed 7 December 2020).

Cornwell S, Victoria R, Niall C, et al. (2016) Fake news or truth? Using satirical cues to detect potentially misleading news. In: *NAACL-CADD 2016: workshop on computational approaches to deception detection at the 15th annual conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp.1–11. San Diego, CA: ACL.

Crowell C (2017) Twitter blog. Available at: https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html (accessed 7 December 2020).

Dale R (2017) NLP in a post-truth world. *Natural Language Engineering* 23(2): 319–324.

Friggeri A, Adamic LA, Eckles D, et al. (2014) Rumor cascades. In: *Proceedings of the eighth international AAAI conference on weblogs and social media*, 2014, pp.101–110. Palo Alto, CA: The AAAI Press.

Garcia-Herranz M, Moro E, Cebrian M, et al. (2014) Using friends as sensors to detect Global-Scale contagious outbreaks. *PLoS ONE* 9(4): e92413.

Gupta A, Kumaraguru P, Castillo C, et al. (2014) Tweetcred: real-time credibility assessment of content on twitter. In: *International conference on social informatics*, 2014, pp. 228–243.

Haigh M, Haigh T and Kozak NI (2018) Stopping fake news. *Journalism Studies* 19(14): 2062–2087.

Hameleers M, Powell TE, Van Der Meer TGLA, et al. (2020) A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication* 37(2): 281–301.

Heverin T and Zach L (2012) Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *Journal of the American Society for Information Science and Technology* 63(1): 34–47.

HLEG on Fake News and Disinformation (2018) A multidimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. Available at: https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation (accessed 7 December 2020).

Hunt K, Wang B and Zhuang J (2020) Misinformation debunking and cross-platform information sharing through Twitter during hurricanes Harvey and Irma: A case study on shelters and ID checks. *Natural Hazards* 103(1): 861–883.

Jacomy M, Venturini T, Heymann S, et al. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE* 9(6): e98679.

Jong W and Dückers MLA (2016) Self-correcting mechanisms and echo-effects in social media: An analysis of the 'gunman in the newsroom' crisis. *Computers in Human Behavior* 59(July): 334–341.

Kwon S, Cha M, Jung K, et al. (2013) Prominent features of rumor propagation in online social media. In: *Proceedings – IEEE international conference on data mining, ICDM*, 2013, pp.1103–1108. New York: IEEE.

Kwon S, Cha M and Jung K (2017) Rumor detection over varying time windows. *Plos ONE* 12(1).

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Mazer JP, Thompson B, Cherry J, et al. (2015) Communication in the face of a school crisis: Examining the volume and content of social media mentions during active shooter incidents. *Computers in Human Behavior* 53: 238–248.

Mirbabaie M and Marx J (2020) 'Breaking' news: Uncovering sense-breaking patterns in social media crisis communication during the 2017 Manchester bombing. *Behaviour & Information Technology* 39(3): 252–266.

Mirbabaie M, Ehnis C, Stieglitz S, et al. (2014) Communication roles in public events: A case study on Twitter communications. In: Doolin B, Lamprou E, Mitev N, et al. (eds.) *Information systems and global assemblages: (re)configuring actors, artefacts, organizations: IFIP WG 8.2 working conference on information systems*, 2014. Heidelberg: Springer.

Newman N, Fletcher R, Levy DAL, et al. (2016) Reuters Institute Digital News Report 2016. Reuters Institute for the Study of Journalism, University of Oxford. Available at: https://reutersinstitute.politics.ox.ac.uk/our-research/digital-news-report-2016 (accessed 7 December 2020).

Nielsen RK, Newman N, Fletcher R, et al. (2019) Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism, University of Oxford. Available at: www.digitalnewsreport.org/survey/2019/foreword-2019/ (accessed 7 December 2020).

Oh O, Agrawal M and Rao HR (2013) Community intelligence and social media services: A rumor theoretic analysis of tweets during social crisis. *MIS Quarterly* 37(2): 407–426.

Pfeffer J, Zorbach T and Carley KM (2014) Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications* 20(1–2): 117–128.

Reporters Without Borders (2020) World Press Freedom Index 2020. Available at: https://rsf.org/en/2020-world-press-freedom-index-entering-decisive-decade-journalism-exacerbated-coronavirus (accessed 7 December 2020).

Ross B, Jung A-K, Heisel J, et al. (2018) Fake news on social media: The (in)effectiveness of warning messages. In: *Proceedings of the 39th international conference on information systems (ICIS)*, San Francisco, CA, USA, 2018.

Roth Y and Pickles N (2020) Updating our approach to misleading information. Available at: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-

to-misleading-information.html (accessed 7 December 2020).

Rubin VL, Chen Y and Conroy NJ (2015) Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Shklovski I, Palen L and Sutton J (2008) Finding community through information and communication technology during disaster events. In: *CSCW '08 proceedings of the 2008 ACM conference on computer supported cooperative work*, 2008, pp. 127–136.

Simon T, Goldberg A, Leykin D, et al. (2016) Kidnapping WhatsApp – rumors during the search and rescue operation of three kidnapped youth. *Computers in Human Behavior* 64: 183–190.

Spiro E, Fitzhugh S, Sutton J, et al. (2012) Rumoring during extreme events: A case study of Deepwater Horizon 2010. In: *Proceeding of the 4th annual ACM web science conference*, 2012, pp.275–283. New York, NY: Association for Computing Machinery.

Spohr D (2017) Fake news and ideological polarization. *Business Information Review* 34(3): 150–160.

Starbird K (2017) Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In: *Proceedings of the 11th international conference on web and social media, ICWSM 2017*, pp.230–239. Palo Alto, CA: The AAAI Press.

Starbird K, Maddock J, Orand M, et al. (2014) Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In: *iConference 2014 proceedings*, 2014, pp.654–662. Grandville, MI: iSchools.

Stieglitz S, Mirbabaie M, Schwenner L, et al. (2017) Sensemaking and communication roles in social media crisis communication. In: *Proceedings der 13. internationalen Tagung Wirtschaftsinformatik (WI 2017)*, St Gallen, Switzerland, 2017, pp.1333–1347.

Stieglitz S, Mirbabaie M, Ross B, et al. (2018) Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management* 39: 156–168.

Surowiecki J (2005) *The Wisdom of the Crowds*. New York: Double Day.

Vargo CJ, Guo L and Amazeen MA (2018) The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society* 20(5): 2028–2049. DOI: 10.1177/1461444817712086.

Vosoughi S, Roy D and Aral A (2018) The spread of true and false news online. *Science* 359(6380): 1146–1151.

Wang B and Zhuang J (2018) Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters. *Natural Hazards* 93(3): 1145–1162.

Wang J, Chen Y, Tang Y, et al. (2016) The effect of rumor clarification on Chinese stock markets. In: *PACIS 2016 proceedings*, 2016, p. 298. Atlanta, GA: Association for Information Systems.

Woodward W (1923) *Bunk*. Manhattan, NY: Harper & Brothers.

Zeng L, Starbird K and Spiro ES (2016) Rumors at the speed of light? Modeling the rate of rumor transmission during crisis. In: *Proceedings of the annual Hawaii international conference on system sciences*, 2016, pp.1969–1978. Washington, DC: IEEE Computer Society.

Zhao Z, Resnick P and Mei Q (2015) Enquiring minds: Early detection of rumors in social media from enquiry posts. In: *Proceedings of the 24th international conference on world wide web*, 2015, pp.1395–1405. New York, NY: Association for Computing Machinery.

Zubiaga A, Liakata M, Procter R, et al. (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS One* 11(3): e0150989.

Zubiaga A, Aker A, Bontcheva K, et al. (2018) Detection and resolution of rumours in social media: a survey. *ACM Computing Surveys 51*(2):32–36.